Temporal-Modal Replay for Continual Audio-Conditioned 3D Motion Generation

Abstract

Audio-conditioned 3D human motion generation models are typically trained once and deployed as static systems, despite real-world audio-motion relationships evolving continuously across speakers and behavioural modes. Naïve sequential fine-tuning catastrophically forgets previously learned motion dynamics, yet continual learning for modern 3D motion generators remains unexplored. We show that forgetting in this setting is governed by *temporal covariance structure* of audio-encoder trajectories. We formalize this through a temporal feature operator and prove that its principal eigenspace uniquely minimizes replay-induced forgetting among equal-rank projections, while whitening is necessary to prevent replay collapse onto dominant motion modes. Motivated by this theory, we propose **Temporal-Modal Replay (TMR)**, the first architecture-agnostic continual learning framework for audio-conditioned 3D human motion generation. TMR preserves temporally informative trajectories through a whitened modal replay representation, requires no auxiliary model or architectural modification, and introduces zero additional trainable parameters or inference overhead. We further introduce the first continual learning benchmark spanning flow matching, autoregressive VQ-VAE, and diffusion-based motion generators across continual speaker and behavioural-mode adaptation. Across twelve baselines, TMR consistently achieves the strongest retention while preserving plasticity, reducing catastrophic forgetting by up to **75%**, achieving lossless generation quality with **25% fewer replay samples**, improving replay feature-space coverage by up to **59%**, and maintaining near-flat long-horizon retention across nine sequential sessions where vanilla fine-tuning catastrophically diverges. For demos, please refer to `https://anony34.github.io/`.

1 Introduction

Audio-conditioned 3D human motion generation aims to synthesize expressive and temporally coherent body motion from auditory signals such as speech, prosody, and spatial acoustics. Recent progress spans three dominant paradigms. In speech-to-gesture generation, autoregressive VQ-VAE models such as TalkSHOW [Yi et al., 2023] generate holistic 3D meshes, while flow-matching frameworks such as GestureLSM [Liu et al., 2025] achieve real-time full-body synthesis. In spatial audio-to-motion generation, diffusion-based methods such as MOSPA [Xu et al., 2025] generate direction-aware motion responses conditioned on binaural audio. Despite architectural differences, these systems share a common structure: an audio encoder produces temporally aligned feature trajectories that condition a downstream autoregressive, flow-matching, or diffusion generator.

Existing methods are trained once on static datasets. In practice, however, audio-motion relationships evolve continuously across two open-world axes: **identity** (new speakers with distinct gesture styles) and **behavioural mode** (different motion responses to similar audio under varying contexts). Retraining large generators from scratch is computationally prohibitive, while naïve fine-tuning causes catastrophic forgetting [Kirkpatrick et al., 2017]. As we show in §3.2, the forgetting magnitude is governed to first order by the gradient interaction $\langle \bar{g}_j, \bar{g}_k \rangle$, which is controlled by the *temporal covariance structure* of encoder trajectories. **No existing continual learning method exploits this structure.**

Preprint.

Current continual learning methods are poorly suited to modern audio-conditioned motion generators. *Regularization-based* methods [Kirkpatrick et al., 2017, Ahn et al., 2019, Eskandar et al., 2025] estimate parameter importance through Fisher or curvature statistics, which become unstable for diffusion, flow-matching, and VQ-VAE objectives. *Replay-based* methods [Tiwari et al., 2022, Tong et al., 2025, Yoon et al., 2021, Aljundi et al., 2019, Van de Ven and Tolias, 2018] reduce each sample to scalar losses or parameter-space gradients, discarding the temporal trajectory structure governing forgetting. *Optimization-based* approaches [Bian et al., 2024, Wu et al., 2024] operate in high-dimensional non-convex parameter landscapes and incur increasing overhead with session count, while *architecture-based* methods [Wu et al., 2025] require pretrained backbones or explicit task boundaries unavailable in our setting. Prior continual gesture generation works [Ahuja et al., 2023, 2022] are restricted to 2D upper-body generation and do not address modern 3D generative architectures. **No prior work studies continual learning for audio-conditioned 3D human motion generation.**

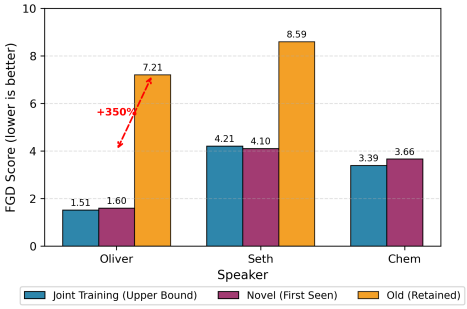

Figure 1: Diagnosing plasticity vs. forgetting. Novel $\approx$ joint (good plasticity), while old $\gg$ novel indicates catastrophic forgetting.

Fig. 1 shows that standard fine-tuning retains strong plasticity, with novel-task Fréchet Gesture Distance (FGD) close to joint training, but catastrophically forgets previous speakers ($\uparrow 350\%$ for Oliver and $\uparrow 210\%$ for Seth). *This suggests that mitigating forgetting, rather than improving plasticity, is the primary challenge.*

We propose **Temporal-Modal Replay (TMR)**, the first architecture-agnostic continual learning framework for audio-conditioned 3D motion generation. We present a unified formulation (§3.1) covering autoregressive, flow-matching, and diffusion generators under two assumptions: **(A1)** audio conditioning occurs through an encoder producing trajectories $\phi_\theta(a) \in \mathbb{R}^{T\times D}$, and **(A2)** the objective is differentiable. Under these assumptions, forgetting is bounded by the *temporal feature operator* $\mathsf{C}_k = \frac{1}{ND}\sum_{i\in\mathcal{D}_{<k}} \tilde{\phi}_i\tilde{\phi}_i^\top \in \mathbb{R}^{T\times T}$, which captures covariance across temporal encoder modes.

We further prove that the top-M eigenspace of C_k uniquely minimizes trajectory reconstruction error among all rank-M projections, yielding the optimal forgetting-minimizing temporal subspace. Whitening the projected trajectories, $\gamma_i = \widehat{\Lambda}_k^{-1/2}\widehat{U}_k^\top \phi_i$, balances dominant and weak temporal modes, preventing replay collapse onto a small subset of trajectories. TMR combines eigenspace estimation, Schwarz-inspired rank selection [Schwarz, 1978], whitening, and sparse greedy recovery into a lightweight replay construction algorithm requiring no auxiliary model, task identity, or architectural modification.

Our contributions are fourfold: **(a)** we formulate continual audio-conditioned 3D motion generation and identify the temporal feature operator C_k as the key structure governing forgetting; **(b)** we develop an optimality theory showing that the principal eigenspace of C_k uniquely minimizes replay forgetting; **(c)** we instantiate this theory in TMR, an efficient architecture-agnostic replay framework with negligible overhead; and **(d)** we introduce the first continual learning benchmark for audio-conditioned 3D motion generation across GestureLSM, TalkSHOW, and MOSPA, where TMR consistently outperforms *twelve* representative regularization, replay, optimization, and architecture-based baselines.

2 Related Work

Audio-conditioned motion generation. Early speech-driven gesture generation methods [Yoon et al., 2020, Ginosar et al., 2019] focused on 2D upper-body motion. TalkSHOW [Yi et al., 2023] extends this to holistic 3D mesh generation using compositional VQ-VAE codebooks and autoregressive decoding. GestureLSM [Liu et al., 2025] employs flow matching with spatial-temporal attention for real-time full-body synthesis on BEAT2 [Liu et al., 2024]. MOSPA [Xu et al., 2025] further introduces spatial audio-conditioned 3D motion generation using diffusion transformers trained on the SAM dataset. These methods represent the dominant autoregressive, flow-matching, and diffusion paradigms for audio-conditioned motion generation, but none addresses continual learning.

Continual learning. Continual learning methods broadly fall into four categories. ***Regularization-based*** approaches such as EWC [Kirkpatrick et al., 2017], UCL [Ahn et al., 2019], and STAR [Eskandar et al., 2025] constrain updates to important parameters using Fisher or uncertainty estimates, which become unstable for stochastic diffusion and flow objectives. ***Replay-based*** methods including ER [Rolnick et al., 2019], MIR [Aljundi et al., 2019], OCS [Yoon et al., 2021], GCR [Tiwari et al., 2022], CSReL [Tong et al., 2025], and generative replay (GenRep) [Van de Ven and Tolias, 2018] store or synthesize previous samples but ignore the temporal covariance structure of encoder trajectories. ***Optimization-based*** methods like C-Flat [Bian et al., 2024], and POCL [Wu et al., 2024] seek update directions compatible with prior tasks, while ***architecture-based*** methods [Wu et al., 2025] (SD-LoRA) expand model capacity to isolate task-specific knowledge. Unlike prior work, TMR directly models the temporal feature covariance C_k that governs forgetting in audio-conditioned motion generation. Detailed per-baseline analyses appear in Appendix H.

Continual gesture generation. DiffGAN [Ahuja et al., 2022] and C-DiffGAN [Ahuja et al., 2023] study continual co-speech gesture generation under speaker adaptation settings. However, they are limited to 2D upper-body gestures and do not address modern 3D motion generators, spatial audio conditioning, or continual behavioural adaptation. To our knowledge, TMR is the first continual learning framework for audio-conditioned 3D human motion generation.

3 Method

3.1 Audio-conditioned 3D motion generation

Let T be the length of a motion clip, d_x the per-frame motion dimension, d_a the per-frame audio dimension, and D the per-frame dimension of the audio feature representation produced by the encoder. A *conditional motion generator* is a parametric map

$$G_\theta : \mathcal{A} \to \mathcal{X}, \qquad \mathcal{X} \subset \mathbb{R}^{T\times d_x}, \;\; \mathcal{A} \subset \mathbb{R}^{T\times d_a}, \;\; \theta \in \mathbb{R}^p, \tag{1}$$

trained by minimising a per-sample loss $\ell(\theta; x, a)$. The three common motion generative models fit Eq. 1 with three distinct losses: a flow-matching velocity-regression loss ℓ^{FM} (flow-matching models [Liu et al., 2025]), a token-level autoregressive log-likelihood ℓ^{AR} (Vector Quantized Variational Autoencoder or VQ-VAE [Yi et al., 2023]), and a denoising diffusion squared-error ℓ^{DM} (diffusion models [Xu et al., 2025]). Their algebraic forms differ; they share three structural features that the present construction exploits.

(A1) *Encoder factorization.* Audio enters the loss through a learned modality encoder $E_{\theta_E} : \mathcal{A} \to \mathbb{R}^{T\times D}$, producing a temporally-aligned representation

$$\phi_\theta(a) \;:=\; \phi_{\theta_E}(a) \;:=\; E_{\theta_E}(a) \;\in\; \mathbb{R}^{T\times D}. \tag{2}$$

The parameters partition as $\theta = (\theta_E, \theta_R)$, with θ_E the encoder block and θ_R the remaining parameters. The per-sample loss factors as

$$\ell(\theta; x, a) = \tilde{\ell}\big(\theta_R,\; \phi_\theta(a),\; x\big), \tag{3}$$

where $\tilde{\ell}$ does not depend directly on θ_E. Thus θ_E influences ℓ exclusively through the audio-conditioned feature trajectory $\phi_\theta(a)$.

(A2) *Differentiability.* ℓ is differentiable in θ, with per-sample gradient $g_i(\theta) := \nabla_\theta \ell(\theta; x_i, a_i) \in \mathbb{R}^p$.

(A3) *Affine gradient structure.* The map $\Phi \mapsto \nabla_\theta \ell(\theta; x, a)\big|_{\phi_{\theta_E}(a)=\Phi}$ is affine in Φ: there exist a matrix A (depending on (x, θ_R)) and a vector b such that $\nabla_\theta \ell\big|_\Phi = A\,\mathrm{vec}(\Phi) + b$. The map is therefore L_∇-Lipschitz with $L_\nabla = \|A\|_2$. This holds, for example, when the decoder $\tilde{\ell}$ is quadratic in Φ or the decoder head is linear; it is used in Proposition 1 to convert trajectory-level approximation errors into gradient-level forgetting bounds.

We refer to any of $\ell^{\mathrm{FM}}, \ell^{\mathrm{AR}}, \ell^{\mathrm{DM}}$ as ℓ when the construction in Section 3.3 treats them uniformly. The encoder output $\phi_\theta(a) \in \mathbb{R}^{T\times D}$ is the fundamental object on which our buffer construction operates: row t is the D-dimensional embedding of the audio context at frame t, and its full $T \times D$ shape is the *audio-conditioned feature trajectory*.

3.2 Continual learning problem and the source of forgetting

Sessions and task losses. The learner observes K ordered sessions. At session $k \in \{1, \ldots, K\}$ a finite sample $\mathcal{D}_k = \{(x_i^{(k)}, a_i^{(k)})\}_{i=1}^{n_k}$ is drawn i.i.d. from a distribution μ_k; a session can correspond

to a new speaker or a new spatial-audio motion (as in MOSPA [Xu et al., 2025]). We write

$$\mathcal{L}_k(\theta) \;:=\; \mathbb{E}_{(x,a)\sim\mu_k}\big[\ell(\theta;x,a)\big], \qquad \bar{g}_k(\theta) \;:=\; \nabla_\theta \mathcal{L}_k(\theta), \tag{4}$$

for the population task loss and its gradient. The ideal joint optimum $\theta_K^\star \in \operatorname*{argmin}_\theta \sum_k \mathcal{L}_k(\theta)$ is unattainable since the learner sees only $\mathcal{D}_k$ at session k.

Catastrophic forgetting. Let θ_{k-1} be the parameters at the end of session $k-1$. Optimising on $\mathcal{D}_k$ with step size $\eta > 0$ yields a displacement of the form

$$\Delta_k := \theta_k - \theta_{k-1} = -\eta\, \bar{g}_k(\theta_{k-1}) + O(\eta^2), \tag{5}$$

where we approximate the empirical update direction by the population gradient $\bar{g}_k$ (accurate for large n_k or in expectation). Taylor-expanding a past task loss $\mathcal{L}_j$ ($j < k$) around θ_{k-1} gives

$$\mathcal{L}_j(\theta_k) - \mathcal{L}_j(\theta_{k-1}) = \langle \bar{g}_j, \Delta_k \rangle + O(\eta^2) = -\eta\, \langle \bar{g}_j, \bar{g}_k \rangle + O(\eta^2). \tag{6}$$

Eq. 6 formalises catastrophic forgetting: past performance degrades when past-task and new-task gradients are positively aligned. Because the loss factorises as $\ell(\theta; x, a) = \tilde{\ell}(\theta_R, \phi_{\theta_E}(a), x)$, the gradients $\bar{g}_j$ and $\bar{g}_k$ depend on the encoder output $\phi_{\theta_E}(a)$. Thus, forgetting on past tasks is influenced by how the audio-conditioned feature trajectories $\phi_{\theta_E}(a_i)$ are distributed across the data.

Replay objective. A bounded replay buffer $\mathcal{B}_{k-1} \subseteq \bigcup_{j<k} \mathcal{D}_j$ with $|\mathcal{B}_{k-1}| \le B$ and weights $w \in \mathbb{R}_{\ge 0}^{|\mathcal{B}_{k-1}|}$ modifies session k's objective to

$$\theta_k \;\in\; \operatorname*{argmin}_\theta \Big\{ \frac{1}{n_k} \sum_{i\in\mathcal{D}_k} \ell(\theta; x_i, a_i) \;+\; \lambda \sum_{i\in\mathcal{B}_{k-1}} w_i\, \ell(\theta; x_i, a_i) \Big\}. \tag{7}$$

The corresponding gradient-descent update (with replay) is

$$\theta_k = \theta_{k-1} - \eta \Big(\bar{g}_k - \lambda \sum_{i\in\mathcal{B}_{k-1}} w_i\, g_i \Big), \tag{8}$$

which is the form analysed in Proposition 1. Substituting into Eq. (5)–(6), forgetting on past task j becomes

$$\mathcal{L}_j(\theta_k) - \mathcal{L}_j(\theta_{k-1}) \;=\; -\eta \left\langle \bar{g}_j,\, \bar{g}_k - \lambda \bar{g}_{\mathcal{B}} \right\rangle + O(\eta^2), \quad \bar{g}_{\mathcal{B}} := \sum_{i\in\mathcal{B}_{k-1}} w_i\, g_i. \tag{9}$$

The buffer's role is therefore to choose $(\mathcal{B}_{k-1}, w)$ so that $\lambda \bar{g}_{\mathcal{B}}$ approximates the projection of $\bar{g}_k$ onto past-task gradient directions. Classical replay does this by matching $\bar{g}_{\mathcal{B}}$ to $\bar{g}_{<k}$ in $\mathbb{R}^p$ [Tiwari et al., 2022, Killamsetty et al., 2021]; this discards the temporal structure $\bar{g}_j$ inherits from $\phi_\theta(a_i)$. We propose to match the buffer to the past distribution *at the level of the audio-conditioned feature trajectories themselves*.

3.3 Temporal-modal replay (TMR)

Audio-conditioned feature trajectories. For a past sample (x_i, a_i) at the previous-session checkpoint θ_{k-1}, define

$$\phi_i \;:=\; \phi_{\theta_{k-1}}(a_i) \;\in\; \mathbb{R}^{T\times D}, \qquad \bar{\phi} \;:=\; \frac{1}{N} \sum_{i\in\mathcal{D}_{<k}} \phi_i, \qquad \tilde{\phi}_i \;:=\; \phi_i - \bar{\phi}, \tag{10}$$

where $N = |\mathcal{D}_{<k}|$. Each ϕ_i is a $T \times D$ matrix: rows are time, columns are encoder feature channels. By property **(A1)**, every per-sample gradient g_i for ℓ factors through ϕ_i, so the distribution of $\{\phi_i\}_{i<k}$ summarises the audio-conditioned behaviour of the model on past data. A coreset that preserves the *distributional geometry* of $\{\phi_i\}$ preserves the audio-driven component of past gradients, and by Eq. 9 preserves past performance.

The temporal feature operator. The geometry of $\{\phi_i\}$ is determined by its second-order statistics. We single out the *temporal* axis because differences in audio-conditioned behaviour over time distinguish samples from the same speaker or spatial-audio motion from those of different ones. Define the *temporal feature operator*

$$\mathsf{C}_k \;:=\; \frac{1}{N\,D} \sum_{i\in\mathcal{D}_{<k}} \tilde{\phi}_i\, \tilde{\phi}_i^\top \;\in\; \mathbb{R}^{T\times T}, \qquad [\mathsf{C}_k]_{t,s} = \frac{1}{ND} \sum_{i,d} \tilde{\phi}_{i,t,d}\, \tilde{\phi}_{i,s,d}. \tag{11}$$

C_k is symmetric positive semi-definite. Its entry $[\mathsf{C}_k]_{t,s}$ is the empirical covariance of the encoder output at time t with time s, averaged across past samples and feature channels; intuitively, it tells us *which time-pairs co-vary across past audio inputs* after the model has been trained.

Diagonalising $\mathsf{C}_k = U_k \Lambda_k U_k^\top$ with eigenvalues $\lambda_k^1 \geq \lambda_k^2 \geq \ldots \geq 0$ and orthonormal columns $u_k^m \in \mathbb{R}^T$. We call u_k^m a *temporal mode*: it is a unit-norm function of frame index t, so projecting a trajectory ϕ_i onto u_k^m yields the D-dimensional coefficient $u_k^{m\,\top} \phi_i \in \mathbb{R}^D$. Trajectories that vary along high-λ modes are the informative ones; trajectories aligned with low-λ modes behave like noise around the mean. Two consequences are important: **(a)** Modes with large λ_k^m span the directions in time along which the past audio-conditioned response varies the most; preserving these modes preserves past behaviour. **(b)** Modes with $\lambda_k^m \to 0$ are noise directions and need not be matched by the buffer.

Theoretical justification (trajectory optimality and forgetting bound). The empirical projector $\widehat{P}_k = \widehat{U}_k \widehat{U}_k^\top$ built from the top $\widehat{M}_k$ modes is not an arbitrary choice.

- **Trajectory reconstruction.** Theorem 1 proves that $\widehat{P}_k$ is the unique minimiser of the empirical trajectory reconstruction error $\Psi(P) = \frac{1}{ND} \sum_i \|(I_T - P)\tilde{\phi}_i\|_F^2$ over all rank-$\widehat{M}_k$ orthogonal projectors (uniqueness follows from the empirical spectral gap, which holds with high probability under the population gap (Assumption 6); see Remark D). Among all possible temporal subspaces of the same dimension, the modal subspace preserves the geometry of the past audio-conditioned trajectories with the smallest possible distortion.
- **Forgetting control.** Proposition 1 shows that, under the affine gradient assumption **(A3)**, the average forgetting gap on past tasks is bounded as

$$\frac{1}{k-1} \sum_{j<k} \big[\mathcal{L}_j(\theta_k) - \mathcal{L}_j(\theta_{k-1})\big] \;\leq\; \eta \, \|\bar{g}_k\|_2 \cdot L_\nabla \, \varepsilon_{\mathcal{B}} \;+\; C\eta^2,$$

  where $\varepsilon_{\mathcal{B}} := \|\frac{1}{N}\sum_i \tilde{\phi}_i - \sum_{i \in \mathcal{B}} w_i \tilde{\phi}_i\|_F$ is the *trajectory approximation error* of the buffer and L_∇ is the Lipschitz constant of the gradient map from **(A3)**. Because $\Psi(\widehat{P}_k)$ is minimised exactly by $\widehat{P}_k$, the modal buffer selection targets precisely this error, making the bound tight.

Modal rank. We choose the truncation rank $\widehat{M}_k$ from the spectrum of C_k via a Schwarz-inspired information criterion (the penalty structure follows Schwarz, 1978),

$$\widehat{M}_k \;=\; \operatorname*{argmin}_{1 \leq M \leq T} \; \log \sigma_M^2 \;+\; \frac{\log N}{T} \, M \, \frac{\sum_{m>M} (\lambda_k^m)^2}{\operatorname{tr}(\mathsf{C}_k)}, \quad \sigma_M^2 := \frac{\operatorname{tr}(\mathsf{C}_k) - \sum_{m \leq M} \lambda_k^m}{T}. \tag{12}$$

Collect the top $\widehat{M}_k$ columns into $\widehat{U}_k = [u_k^1, \ldots, u_k^{\widehat{M}_k}] \in \mathbb{R}^{T \times \widehat{M}_k}$ and the corresponding eigenvalues into $\widehat{\Lambda}_k = \operatorname{diag}(\lambda_k^1, \ldots, \lambda_k^{\widehat{M}_k})$.

Statistical consistency. Theorem 2 provides a rigorous finite-sample analysis: under bounded encoder outputs, sub-Gaussian entries, effective rank $r^* = \operatorname{tr}(\mathsf{C}_k^*)/\|\mathsf{C}_k^*\|_2$, and the sample-size condition $N \gtrsim T^2/(r^* \log T)$, the empirical projector $\widehat{P}_k$ and the trajectory error $\Psi(\widehat{P}_k)$ converge to their population counterparts at rate $O\big(\sqrt{r^* \log T / N}\big)$ with probability at least $1 - 5T^{-2}$. The effective-rank factor r^* reflects the intrinsic complexity of the temporal covariance and matches the minimax lower bound for covariance estimation [Koltchinskii and Lounici, 2017]. This guarantees that the eigendecomposition performed on the finite buffer faithfully recovers the true population temporal modes, and the optimality guarantees from Theorem 1 and Proposition 1 extend to the full data distribution.

Whitened modal representation. Project each past sample into the dominant temporal subspace and whiten by $\widehat{\Lambda}_k^{-1/2}$:

$$\beta_i \;:=\; \widehat{U}_k^\top \phi_i \;\in\; \mathbb{R}^{\widehat{M}_k \times D}, \qquad \gamma_i \;:=\; \widehat{\Lambda}_k^{-1/2} \, \beta_i \;\in\; \mathbb{R}^{\widehat{M}_k \times D}. \tag{13}$$

The matrix γ_i is the *whitened modal representation* of sample i: it records the energy of ϕ_i along the dominant temporal modes, equalised across modes by whitening so that no single mode dominates Euclidean inner products. Whitening is essential here, as without it, the largest mode would absorb almost all of the signal, and the coreset selector would collapse to picking samples that align with u_k^1 alone.

Selection rule. Vectorise the representation, $z_i := \mathrm{vec}(\gamma_i)/\|\mathrm{vec}(\gamma_i)\|_2 \in \mathbb{R}^{\widehat{M}_k D}$ (unit-norm), and form the design matrix $Z_k := [z_1, \ldots, z_N]$ with target $y_k := \sum_{i=1}^{N} z_i$. The optimal replay buffer $(\mathcal{B}_k, w^\star)$ is the solution of

$$(\mathcal{B}_k, w^\star) = \underset{\substack{\mathcal{B} \subseteq \mathcal{D}_{<k},\, |\mathcal{B}| \le B \\ w \in \mathbb{R}^{|\mathcal{B}|}_{\ge 0}}}{\mathrm{argmin}} \|y_k - \sum_{i \in \mathcal{B}} w_i z_i\|_2^2. \tag{14}$$

Eq. 14 formulates the selection of a sparse, non-negatively weighted subset of past samples, where modal representations approximate the population target y_k via a ridge-regularized non-negative least-squares objective. It is a sparse non-negative recovery problem on the dictionary Z_k with target y_k. We solve it by an iterative greedy-support algorithm (Algorithm 1 in appendix) that, at each iteration, augments the active support by the indices of largest non-negative correlation with the current residual, refits non-negative ridge-regularized weights and updates the residual. After B samples have been selected, the weights are normalised to mean unity, $w_i \leftarrow B w_i / \sum_j w_j$, and $(\mathcal{B}_k, w^\star)$ is passed to the replay objective (Eq. 7).

Compressive Sampling Matching Pursuit (CoSaMP) [Needell and Tropp, 2009] identifies a proxy support $\Omega = \mathrm{supp}_{2B}(Z_k^\top r)$, merges it with the current support to $\mathcal{T} = \mathcal{S} \cup \Omega$, solves an unconstrained least-squares $\tilde{w} = \mathrm{argmin}_w \|y_k - Z_{\mathcal{T}} w\|_2^2$, and prunes to the B largest entries by magnitude, $\mathcal{S} \leftarrow \mathrm{supp}_B(\tilde{w})$. **Subspace Pursuit (SSP)** [Dai and Milenkovic, 2009] similarly builds $\mathcal{T} = \mathcal{S} \cup \Omega$ of size at most $2B$, computes $\tilde{w}$ on $\mathcal{T}$, prunes to the B largest entries $\mathcal{S}' \leftarrow \mathrm{supp}_B(\tilde{w})$, and then re-solves the least-squares on $\mathcal{S}'$ to obtain the new residual r'. It terminates if $\|r'\|_2 > \|r\|_2$, reverting to the previous iterate; otherwise it proceeds. **Orthogonal Matching Pursuit (OMP)** [Tropp and Gilbert, 2007] adds one atom per iteration via maximum non-negative correlation with the residual, $i^\ell = \mathrm{argmax}_{i \notin \mathcal{S}} \langle z_i, r^{\ell-1} \rangle_+$, and then refits all weights on the augmented support by solving the non-negative least-squares $\min_{w \ge 0} \|y_k - Z_{\mathcal{S} \cup \{i^\ell\}} w\|_2^2$. We evaluate three greedy pursuit algorithms to solve the non-negative sparse recovery problem (14).

4 Implementation and Dataset Details

Experimental Setup. We adopt a session-wise continual learning protocol across all benchmarks. Session 0 trains the model from scratch on the first speaker or genre, while each subsequent session sequentially introduces a new speaker or motion genre without access to prior raw data. To mitigate catastrophic forgetting, we employ Temporal-Modal Replay (TMR), constructing a replay buffer containing 10% of each session's data via Orthogonal Matching Pursuit (OMP) for non-negative sparse recovery, with CoSaMP and SSP as alternative solvers (Section 3.3). Replay buffers accumulate compressed representations of all previously observed sessions. Code will be released upon acceptance. All experiments are performed on an A100 40GB GPU.

Table 1: **Continual speaker learning on BEAT2 using FGD (↓).** TMR achieves the lowest forgetting (F^{avg}) while maintaining adaptation performance comparable to all baselines on flow-based models.

Session No.	Speaker	Test	Vanilla	TMR	ER	EWC	UCL	OCS	STAR	C-Flat	CSReL	GenRep	MIR	POCL	GCR
0	2	2	**0.412**												
1	3	2	1.489	**0.454**	0.618	1.540	1.515	0.683	0.485	0.664	0.558	0.567	0.550	0.676	0.486
		3	2.260	2.229	2.249	2.380	2.250	2.200	2.330	2.234	2.250	2.616	2.170	2.124	2.264
2	11	2	2.140	**0.488**	0.632	2.160	2.091	0.650	0.923	0.718	0.534	1.886	0.603	0.804	0.651
		3	3.160	**2.483**	2.830	3.150	2.986	2.653	3.045	3.187	2.504	2.954	2.489	2.574	2.538
		11	2.059	2.078	2.080	2.780	2.274	2.106	2.195	2.014	2.043	1.988	2.065	2.069	2.076
3	7	2	1.874	**0.604**	0.815	2.346	2.183	0.726	0.904	0.948	0.658	2.194	0.684	0.985	0.675
		3	3.241	**2.551**	2.591	3.374	3.109	2.894	3.163	3.163	2.819	2.962	2.805	2.648	2.602
		11	2.964	**2.126**	3.086	2.815	3.015	3.301	2.879	2.847	2.903	2.378	2.313	2.700	2.814
		7	3.508	3.571	3.515	3.626	3.569	3.610	3.522	3.675	3.681	3.956	3.710	3.640	3.670
4	1	2	2.129	**0.613**	0.880	2.327	2.191	1.146	1.389	1.126	0.743	2.347	0.977	1.146	0.732
		3	3.215	2.605	2.718	3.489	3.164	2.801	2.674	2.761	2.589	2.880	2.741	2.871	**2.571**
		11	2.972	**2.624**	3.114	2.980	3.210	3.027	3.010	2.942	2.740	2.631	2.758	2.635	2.862
		7	4.112	**3.570**	3.894	3.750	3.626	3.781	3.640	3.675	3.814	3.620	3.915	3.610	3.895
		1	2.392	2.565	2.630	2.610	2.681	2.656	2.791	2.710	2.629	2.641	2.610	2.634	2.632
		F^{avg} ↓	0.698	**0.185**	0.454	0.551	0.584	0.463	0.583	0.476	0.312	0.551	0.362	0.397	0.356
		A^{avg} ↓	**2.555**	2.611	2.619	2.849	2.694	2.643	2.709	2.658	2.651	2.800	2.639	2.617	2.660

Datasets and Benchmarks. We evaluate on three benchmarks spanning distinct motion representations, generator families, and continual distribution shifts. *(i) BEAT2 with GestureLSM* uses flow

matching for co-speech gesture generation across nine speakers (Wayne, Scott, Solomon, Lawrence, Carla, Sophie, Miranda, Nidal, Ayana), forming a rigorous nine-session continual speaker-learning benchmark (Table 4, appendix). *(ii) TalkSHOW with VQ-VAE* performs holistic 3D mesh generation conditioned on speech over three speakers (Oliver, Seth, Chemistry), where continual fine-tuning risks overwriting discrete gesture codebooks (Table 5, appendix). *(iii) SAM with MOSPA* employs a diffusion backbone and introduces continual shift through motion genre: *Insensitive* $\rightarrow$ *Sensitive* $\rightarrow$ *Neutral*, forming a deliberate worst-case forgetting curriculum with progressively contrasting motion responses (Table 6, appendix). **Metrics.** We report Fréchet Inception Distance (FID) [Heusel et al., 2017] and Fréchet Gesture Distance (FGD) [Yoon et al., 2020], which measure distributional similarity between generated and real motions in learned feature space and are highly sensitive to mode collapse and distributional drift. We further report temporal-average forgetting and adaptation error: $F_i^{avg} = \frac{1}{T-i}\sum_{t=i+1}^{T}(m_{i,t} - m_{i,i}), \quad F^{avg} = \frac{1}{|\mathcal{S}_{old}|}\sum_i F_i^{avg}, \quad A^{avg} = \frac{1}{|\mathcal{S}_{new}|}\sum_i m_{i,i}$. Lower values are better for both forgetting and adaptation error. Across all benchmarks, forgetting rather than plasticity is the dominant continual-learning failure mode. **Hyperparameters.** TMR introduces virtually no hyperparameters. Replay size is fixed at 10% of each session with replay weight λ=1, while the modal rank $\widehat{M}_k$ is estimated automatically from the data. All remaining settings follow the default backbone configurations.

5 Results and Analysis

We evaluate TMR on three continual audio-conditioned 3D motion generation benchmarks: GestureLSM on BEAT2, TalkSHOW on holistic 3D meshes, and MOSPA on SAM. Tables 1, 2, and 3 compare against replay-, regularization-, optimization-, and architecture-based continual learning baselines under identical replay budgets and training conditions. A highly consistent trend emerges across all experiments: **plasticity is already strong for nearly all methods, whereas forgetting dominates the performance gap.** Vanilla fine-tuning typically achieves adaptation errors (A^{avg}) close to continual-learning methods, but catastrophically destroys old speaker/motion representations over sequential sessions. TMR specifically targets this failure mode and consistently achieves the lowest F^{avg} across all benchmarks.

Table 2: **Continual 3D mesh generation on TalkSHOW using FGD (↓) with VQ-VAE.**

Session No.	Speaker	Vanilla	TMR	TMR (CoSaMP)	TMR (SSP)	GCR	MIR	CSReL	OCS	ER
0	Oliver	**1.600**								
1	Oliver	6.193	**1.982**	2.121	2.015	2.558	2.343	2.321	2.569	2.419
	Seth	4.100	4.267	4.474	4.571	4.751	4.403	4.754	4.973	4.872
2	Oliver	8.220	2.314	2.346	**2.301**	3.369	2.828	2.427	3.031	3.410
	Seth	8.590	**4.411**	4.426	4.457	4.722	4.456	4.488	4.895	4.983
	Chem	3.658	4.080	4.102	4.114	3.676	4.163	4.033	5.088	5.017
	F^{avg} ↓	5.048	0.346	0.293	**0.222**	0.667	0.519	0.254	0.561	0.713
	A^{avg} ↓	**3.658**	4.080	4.102	4.114	3.676	4.163	4.033	5.088	5.017

Table 3: **Continual motion-type learning on MOSPA using FID (↓) with Diffusion.**

Session No.	Genre	Vanilla	TMR	OCS	MIR	GCR	CSReL
0	Insensitive	**26.503**					
1	Insensitive	161.899	**39.157**	72.065	70.160	47.132	125.228
	Sensitive	19.457	19.635	19.875	19.943	19.828	22.389
2	Insensitive	61.477	**42.178**	47.354	53.110	51.538	42.993
	Sensitive	62.617	**37.802**	43.700	39.860	40.144	67.617
	Neutral	19.990	25.997	23.297	24.690	23.900	21.800
	F^{avg} ↓	64.173	**16.166**	28.516	27.524	21.574	51.418
	A^{avg} ↓	**19.724**	22.816	21.586	22.317	21.864	22.095

On GestureLSM (Table 1), TMR achieves substantially lower F^{avg} than MIR, GCR, OCS, ER, and CSReL while maintaining nearly identical current-session FGD, confirming that replay quality—not adaptation capacity—is the key factor. A striking example is Speaker 2: vanilla fine-tuning drifts from 0.412 to 2.129 FGD after sequential updates ($+417\%$), whereas TMR retains it at 0.613 ($+49\%$), corresponding to a $\sim 3.4\times$ smaller degradation. Importantly, TMR's advantage compounds with

session count as the temporal feature operator C_k accumulates richer cross-speaker covariance structure over time. We noted that the performance of SD-LoRA is similar to that of EWC and UCL.

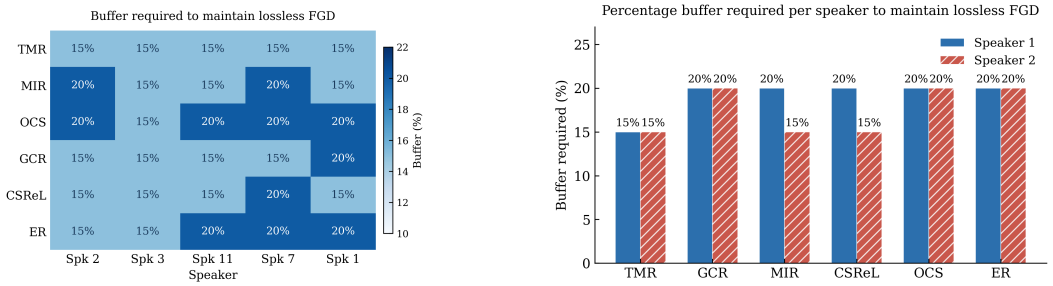

Figure 2: **(a)** Minimum replay percentage for lossless FGD retention on GestureLSM. **(b)** Minimum replay percentage for lossless FGD retention on TalkSHOW. TMR achieves lossless retention with smaller replay buffers.

The same phenomenon appears on TalkSHOW (Table 2) despite the fundamentally different autoregressive VQ-VAE architecture. Vanilla fine-tuning catastrophically overwrites earlier gesture-token distributions: Oliver deteriorates from 1.600 to 8.220 FGD (+414%), whereas TMR preserves Oliver near the original quality regime (2.314 under OMP and 2.301 under SSP). SSP and CoSaMP further improve forgetting, showing that the sparse recovery stage is modular while the modal replay formulation remains stable. Current-speaker FGD remains tightly clustered across methods, again confirming that forgetting—not adaptation—is the dominant challenge. Among non-TMR methods, CSReL performs strongest but still underperforms TMR while requiring substantially larger replay-construction overhead.

On MOSPA (Table 3), the continual axis shifts from speaker identity to motion behavior, producing the most difficult distributional transition in the paper. Vanilla fine-tuning reaches extremely severe forgetting (F^{avg}=64.173), particularly during the Insensitive→Sensitive transition where the old and new motion manifolds are nearly orthogonal. Insensitive FID explodes from 26.503 to 161.899 under vanilla (+511%), whereas TMR limits this to 39.157 (+47.7%). By Session 2, vanilla oscillates between recently abandoned genres, revealing classic catastrophic interference, while TMR simultaneously preserves both old and recent motion distributions with substantially lower FID across prior genres. *This highlights that TMR generalizes beyond speaker identity and remains effective under fundamentally different behavioral-mode shifts and diffusion-based generators.*

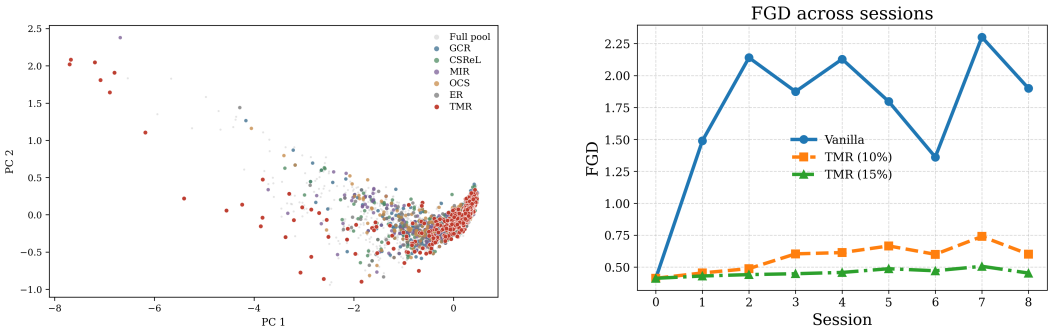

Figure 3: **(a)** Temporal replay coverage in feature space. **(b)** Long-horizon forgetting on GestureLSM (BEAT2).

Long-horizon continual learning stress test. Figure 3b and Appendix Table 8 evaluate a nine-session continual-learning stress test on GestureLSM. Vanilla forgetting compounds monotonically, with Speaker 2 deteriorating to 1.900 FGD by Session 8 (+361% from Session 0). Using only a 15% replay buffer, TMR maintains Speaker 2 near-flat across all sessions, ending at **0.601**, **remarkably close to the original Session 0 value of** 0.412. Importantly, TMR fluctuates only minimally between Sessions 4 and 8, whereas vanilla continuously drifts upward, indicating that forgetting does not compound even as the replay buffer must represent increasingly heterogeneous speakers. This directly aligns with the theory in Section 3.3: as sessions accumulate, C_k captures progressively richer

cross-speaker temporal covariance structure, making the modal replay representation increasingly representative rather than weaker.

Replay efficiency and lossless retention. Figures 2a and 2b evaluate replay efficiency through the minimum replay percentage required to achieve *lossless* FGD, i.e., continual performance matching the joint-training upper bound. TMR consistently achieves lossless retention with only 15% replay across both benchmarks and all speakers, whereas competing methods require at least 20% for multiple speakers, corresponding to roughly a 25% reduction in replay memory. This is particularly important since replay memory scales linearly with session count in long-horizon continual learning. Importantly, competing methods exhibit strong speaker-to-speaker variability in required replay percentage, indicating unstable replay selection under heterogeneous motion distributions, whereas TMR remains uniformly stable across speakers, suggesting that the **modal replay representation adapts robustly to differing motion styles and audio-motion correlations.**

Feature-space coverage and why TMR works. Figure 3a explains the superior replay quality of TMR. We project temporal audio-feature trajectories for Speaker 2 into PCA space and evaluate replay coverage using $\mathrm{cov} = \frac{|\,\mathcal{C}_{\text{buf}} \cap \mathcal{C}_{\text{full}}\,|}{|\,\mathcal{C}_{\text{full}}\,|} \in [0, 1]$, where $\mathcal{C}_{\text{full}}$ and $\mathcal{C}_{\text{buf}}$ denote occupied grid cells in the full feature distribution (Full pool) and replay buffer, respectively. TMR achieves $\mathrm{cov}{=}0.481$, outperforming MIR (0.406) by 18.5% relative and ER (0.302) by nearly 59%. Existing replay methods cluster around dense central regions of the speaker manifold, preserving mainly majority motion behavior, whereas TMR allocates replay samples into sparse outer regions corresponding to rare but highly discriminative trajectories that receive little future gradient signal and are therefore forgotten first under vanilla fine-tuning. The whitening step, $\gamma_i = \widehat{\Lambda}_k^{-1/2} \widehat{U}_k^\top \phi_i$, suppresses dominant high-density modes and amplifies weaker temporal directions, preventing replay collapse onto only the most common trajectories. Consequently, *TMR preserves both dominant and low-density temporal structures, yielding broader feature-space coverage and lower forgetting.*

Computational complexity and practical overhead. Table 7 in Appendix, shows that TMR achieves these gains with negligible overhead. Replay construction remains comparable to or faster than existing replay baselines: TMR constructs its buffer in $1{:}46$ mins, substantially faster than GCR ($3{:}31$ mins) and CSReL ($4{:}29$ mins), while requiring modest VRAM (2028 MiB), comparable to GCR and substantially lower than CSReL (5350 MiB). Once replay construction is complete, all methods train under identical conditions with the same VRAM, training iterations, replay size, and model capacity. *Thus TMR achieves substantially lower forgetting without increasing training or inference complexity and introduces zero additional trainable parameters.*

Overall, the results validate the central claim of this work: **continual audio-conditioned motion generation is fundamentally governed by preservation of temporal covariance structure rather than replay quantity or parameter-space similarity.** Across flow matching, autoregressive VQ-VAE, and diffusion generators, TMR preserves this structure substantially more effectively than existing replay methods, yielding consistently lower forgetting, stronger long-horizon stability, broader feature-space coverage, superior replay efficiency, and negligible computational overhead. Please refer to Appendix I for a detailed discussion on Limitations and Future work.

6 Conclusion

We introduced the first continual learning framework for modern audio-conditioned 3D human motion generation. We showed that forgetting is fundamentally governed by the temporal covariance structure of audio encoder trajectories and proposed Temporal-Modal Replay (TMR), an architecture-agnostic replay framework based on whitened temporal-modal representations. Across flow matching, autoregressive VQ-VAE, and diffusion generators, TMR consistently achieves substantially lower forgetting, stronger long-horizon retention, broader feature-space coverage, and higher replay efficiency than existing continual learning methods while introducing zero additional inference overhead or trainable parameters. These results establish temporal-modal replay as an effective and scalable foundation for continual audio-conditioned 3D motion generation.

References

Hongjoon Ahn, Sungmin Cha, Donggyu Lee, and Taesup Moon. Uncertainty-based continual learning with adaptive regularization. *Advances in neural information processing systems*, 32, 2019.

Chaitanya Ahuja, Dong Won Lee, and Louis-Philippe Morency. Low-resource adaptation for personalized co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20566–20576, 2022.

Chaitanya Ahuja, Pratik Joshi, Ryo Ishii, and Louis-Philippe Morency. Continual learning for personalized co-speech gesture generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20893–20903, 2023.

Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. Online continual learning with maximal interfered retrieval. *Advances in neural information processing systems*, 32, 2019.

Ang Bian, Wei Li, Hangjie Yuan, Chengrong Yu, Mang Wang, Zixiang Zhao, Aojun Lu, Pengliang Ji, and Tao Feng. Make continual learning stronger via c-flat. *Advances in Neural Information Processing Systems*, 37:7608–7630, 2024.

Wei Dai and Olgica Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE transactions on Information Theory*, 55(5):2230–2249, 2009.

Masih Eskandar, Tooba Imtiaz, Davin Hill, Zifeng Wang, and Jennifer Dy. Star: Stability-inducing weight perturbation for continual learning. *arXiv preprint arXiv:2503.01595*, 2025.

Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3497–3506, 2019.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pages 5464–5474. PMLR, 2021.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pages 110–133, 2017.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1144–1154, 2024.

Pinxin Liu, Luchuan Song, Junhua Huang, Haiyang Liu, and Chenliang Xu. Gesturelsm: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10929–10939, 2025.

Deanna Needell and Joel A Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and computational harmonic analysis*, 26(3):301–321, 2009.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

Gilbert W Stewart and Ji-guang Sun. Matrix perturbation theory. *(No Title)*, 1990.

Rishabh Tiwari, Krishnateja Killamsetty, Rishabh Iyer, and Pradeep Shenoy. Gcr: Gradient coreset based replay buffer selection for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 99–108, 2022.

Ruilin Tong, Yuhang Liu, Javen Qinfeng Shi, and Dong Gong. Coreset selection via reducible loss in continual learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666, 2007.

GM Van de Ven and AS Tolias. Generative replay with feedback connections as a general strategy for continual learning. arxiv. *arXiv preprint arXiv:1809.10635*, 2018.

Yichen Wu, Hong Wang, Peilin Zhao, Yefeng Zheng, Ying Wei, and Long-Kai Huang. Mitigating catastrophic forgetting in online continual learning by modeling previous task interrelations via pareto optimization. In *Forty-first international conference on machine learning*, 2024.

Yichen Wu, Hongming Piao, Long-Kai Huang, Renzhen Wang, Wanhua Li, Hanspeter Pfister, Deyu Meng, Kede Ma, and Ying Wei. Sd-lora: Scalable decoupled low-rank adaptation for class incremental learning. *arXiv preprint arXiv:2501.13198*, 2025.

Shuyang Xu, Zhiyang Dou, Mingyi Shi, Liang Pan, Leo Ho, Jingbo Wang, Yuan Liu, Cheng Lin, Yuexin Ma, Wenping Wang, et al. Mospa: Human motion generation driven by spatial audio. *arXiv preprint arXiv:2507.11949*, 2025.

Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 469–480, 2023.

Jaehong Yoon, Divyam Madaan, Eunho Yang, and Sung Ju Hwang. Online coreset selection for rehearsal-based continual learning. *arXiv preprint arXiv:2106.01085*, 2021.

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020.

Appendix

A Algorithm

Algorithm 1 Temporal-Modal Replay (TMR)

Require: Past samples $\mathcal{D}_{<k}$, checkpoint θ_{k-1}, budget B.

Compute $\phi_i = \phi_{\theta_{k-1}}(a_i)$ for each $i \in \mathcal{D}_{<k}$ and centre to obtain $\tilde{\phi}_i$ (Eq. 10).
Form C_k via Eq. 11; eigendecompose $\mathsf{C}_k = U_k \Lambda_k U_k^\top$.
Choose $\widehat{M}_k$ by Eq. 12; extract $(\widehat{U}_k, \widehat{\Lambda}_k)$.
Compute γ_i via Eq. 13; flatten and ℓ_2-normalise to obtain unit vectors z_i.
Initialise residual $r \leftarrow y_k$ and active support $\mathcal{S} \leftarrow \emptyset$.
repeat
  Augment. Select indices $\mathcal{T} \subseteq \{1, \dots, N\} \setminus \mathcal{S}$ of largest non-negative correlation $\langle z_i, r \rangle_+$ with the residual;
  Update $\mathcal{S} \leftarrow \mathcal{S} \cup \mathcal{T}$.
  Refit. Solve $w_\mathcal{S} \leftarrow \underset{w \geq 0}{\operatorname{argmin}} \|y_k - Z_\mathcal{S} w\|_2^2$;
  Update residual $r \leftarrow y_k - Z_\mathcal{S} w_\mathcal{S}$.
until $|\mathcal{S}| = B$
Normalise weights $w_i \leftarrow B w_i / \sum_j w_j$.

Ensure: Replay buffer $(\mathcal{B}_k, w^\star) = (\mathcal{S}, w_\mathcal{S})$ used in Eq. 7.

B Setup and Notation

- T: number of frames; D: encoder feature dimension; $N = |\mathcal{D}_{<k}|$: total past samples.
- $\phi_i \in \mathbb{R}^{T \times D}$: audio-conditioned feature trajectory of sample i at checkpoint θ_{k-1}.
- $\bar{\phi}^* := \mathbb{E}_\mu[\phi]$: population mean; $\bar{\phi} := \frac{1}{N} \sum_i \phi_i$: empirical mean; $\delta := \bar{\phi} - \bar{\phi}^*$.
- $\tilde{\phi}_i^* := \phi_i - \bar{\phi}^*$ (population-centred, used in analysis); $\tilde{\phi}_i := \phi_i - \bar{\phi}$ (sample-centred, used in algorithm). Note $\tilde{\phi}_i = \tilde{\phi}_i^* - \delta$.
- Population temporal operator: $C_k^* := \mathbb{E}_\mu \left[\frac{1}{D} \tilde{\phi}^* (\tilde{\phi}^*)^\top \right] \in \mathbb{R}^{T \times T}$.
- Empirical temporal operator: $\mathsf{C}_k := \frac{1}{ND} \sum_{i=1}^N \tilde{\phi}_i \tilde{\phi}_i^\top \in \mathbb{R}^{T \times T}$.
- Eigenvalues: $\lambda_k^1 \geq \cdots \geq \lambda_k^T \geq 0$ for C_k; $\lambda^{*,1} \geq \cdots \geq \lambda^{*,T} \geq 0$ for C_k^*.
- $\widehat{P}_k = \widehat{U}_k \widehat{U}_k^\top$: empirical rank-$\widehat{M}_k$ projector onto top eigenvectors of C_k; $P_k^* = U_k^* U_k^{*\top}$: population rank-$M^*$ projector.
- Effective rank: $r^* := \operatorname{tr}(C_k^*) / \|C_k^*\|_2 \geq 1$.
- $X_i^* := \frac{1}{D} \tilde{\phi}_i^* (\tilde{\phi}_i^*)^\top, \quad S_N := \frac{1}{N} \sum_{i=1}^N X_i^*, \quad$ so $C_k^* = \mathbb{E}[X_i^*]$.

C Assumptions

Assumption 1 (Encoder factorisation)**.** *$\ell(\theta; x, a) = \tilde{\ell}(\theta_R, \phi_{\theta_E}(a), x)$ with $\tilde{\ell}$ independent of θ_E directly. Per-sample gradients factor through ϕ_i via the chain rule.*

Assumption 2 (Bounded encoder output)**.** *$\|\phi_i\|_2 \leq B_\phi$ almost surely. Hence $\left\|\tilde{\phi}_i^*\right\|_2 \leq 2B_\phi$ and, by the rank-1 identity $\left\|vv^\top\right\|_2 = \|v\|_2^2$:*

$$\|X_i^*\|_2 = \frac{\left\|\tilde{\phi}_i^*\right\|_2^2}{D} \leq \frac{4B_\phi^2}{D} =: R \quad \textit{almost surely.} \tag{15}$$

Assumption 3 (End-to-end gradient affine)**.** *The map $\Phi \mapsto \nabla_\theta \ell(\theta; x, a)\big|_{\phi_{\theta_E}(a) = \Phi}$ is affine: there exists a matrix A (depending on (x, θ_R)) and a vector b such that for any Φ,*

$$\nabla_\theta \ell\big|_\Phi = A \operatorname{vec}(\Phi) + b.$$

Consequently, the map is L_∇-Lipschitz with $L_\nabla = \|A\|_2$, and moreover for any two convex combinations $\sum_i \alpha_i \Phi_i$ and $\sum_i \beta_i \Phi_i$,

$$\Big\|\sum_i \alpha_i g(\Phi_i) - \sum_i \beta_i g(\Phi_i)\Big\|_2 \;\le\; L_\nabla \Big\|\sum_i \alpha_i \Phi_i - \sum_i \beta_i \Phi_i\Big\|_F.$$

This holds, for example, when the decoder is linear or the loss is quadratic.

Assumption 4 (Sub-Gaussian entries)**.** *Each scalar entry of $\tilde{\phi}_i^*$ is σ-sub-Gaussian: $\mathbb{E}[e^{s[\tilde{\phi}_i^*]_{t,d}}] \le e^{s^2\sigma^2/2}$ for all $s \in \mathbb{R}$.*

Assumption 5 (Bounded effective rank)**.** $r^* := \mathrm{tr}(C_k^*)/\|C_k^*\|_2 \le r_{\max}$ *with* $r_{\max} = o(N/\log T)$.

Assumption 6 (Spectral gap)**.** $\delta_{\mathrm{gap}} := \lambda^{*,M^*} - \lambda^{*,M^*+1} > 0$.

D Theorem 1: Trajectory Optimality

Theorem 1 (Trajectory optimality)**.** *Let $\mathcal{P}_M$ be the set of rank-M orthogonal projectors in $\mathbb{R}^{T\times T}$. Define $\Psi(P) := \frac{1}{ND}\sum_{i=1}^N \Big\|(I_T - P)\tilde{\phi}_i\Big\|_F^2$. Then:*

$$\widehat{P}_k \;=\; \operatorname*{argmin}_{P\in\mathcal{P}_{\widehat{M}_k}} \Psi(P), \qquad \Psi(\widehat{P}_k) \;=\; \sum_{m>\widehat{M}_k} \lambda_k^m.$$

The minimiser is unique whenever $\lambda_k^{\widehat{M}_k} > \lambda_k^{\widehat{M}_k+1}$; this empirical gap holds with high probability by Theorem 2 and Weyl's inequality (see Remark D).

Proof. Since P is an orthogonal projector, $(I_T - P)^2 = I_T - P$. Therefore:

$$\Psi(P) = \frac{1}{ND}\sum_i \mathrm{tr}\big(\tilde{\phi}_i^\top (I_T - P)\tilde{\phi}_i\big) = \mathrm{tr}\big((I_T - P)\,\mathsf{C}_k\big) = \mathrm{tr}(\mathsf{C}_k) - \mathrm{tr}(P\mathsf{C}_k).$$

Minimising $\Psi(P)$ is equivalent to maximising $\mathrm{tr}(P\mathsf{C}_k)$. Write $P = VV^\top$ with $V \in \mathbb{R}^{T\times\widehat{M}_k}$, $V^\top V = I$; then $\mathrm{tr}(P\mathsf{C}_k) = \mathrm{tr}(V^\top \mathsf{C}_k V)$. By the Courant–Fischer theorem, $\max_V \mathrm{tr}(V^\top\mathsf{C}_k V) = \sum_{m=1}^{\widehat{M}_k}\lambda_k^m$, achieved when V spans the top-$\widehat{M}_k$ eigenspace of C_k, i.e. $V = \widehat{U}_k$. Therefore $\widehat{P}_k$ is the minimiser and $\Psi(\widehat{P}_k) = \mathrm{tr}(\mathsf{C}_k) - \sum_{m=1}^{\widehat{M}_k}\lambda_k^m = \sum_{m>\widehat{M}_k}\lambda_k^m$. □

Remark (Uniqueness)**.** *Assumption 6 gives $\delta_{\mathrm{gap}} > 0$ at the population level. Theorem 2 (proved independently of Theorem 1) yields $\|\mathsf{C}_k - C_k^*\|_2 \le \varepsilon$ with high probability. By Weyl's inequality, $|\lambda_k^m - \lambda^{*,m}| \le \varepsilon$ for every m. When $\varepsilon < \delta_{\mathrm{gap}}/2$, the empirical gap $\lambda_k^{\widehat{M}_k} - \lambda_k^{\widehat{M}_k+1} \ge \delta_{\mathrm{gap}} - 2\varepsilon > 0$, so the top-$\widehat{M}_k$ eigenspace of C_k is unique and $\widehat{P}_k$ is the unique minimiser. The required N is given by condition* (20)*.*

Proposition 1 (Forgetting reduction)**.** *Under Assumptions 1 and 3, let $(\mathcal{B}, w)$ be any replay buffer with $w_i \ge 0$ and $\sum_{i\in\mathcal{B}} w_i = 1$. Define the* trajectory approximation error*:*

$$\varepsilon_{\mathcal{B}} \;:=\; \left\|\frac{1}{N}\sum_{i=1}^N \tilde{\phi}_i - \sum_{i\in\mathcal{B}} w_i\,\tilde{\phi}_i\right\|_F, \tag{16}$$

where $\tilde{\phi}_i = \phi_i - \bar{\phi}$ are the centred trajectories. Suppose the parameter update is

$$\theta_k = \theta_{k-1} - \eta\Big(\bar{g}_k - \sum_{i\in\mathcal{B}} w_i g_i\Big),$$

where $\bar{g}_k = \frac{1}{N}\sum_{i=1}^N \nabla_\theta \ell(\theta_{k-1}; x_i, a_i)$. Then there exists an absolute constant $C > 0$ such that

$$\frac{1}{k-1}\sum_{j<k}\big[\mathcal{L}_j(\theta_k) - \mathcal{L}_j(\theta_{k-1})\big] \;\le\; \eta\,\|\bar{g}_k\|_2 \cdot L_\nabla\,\varepsilon_{\mathcal{B}} \;+\; C\eta^2. \tag{17}$$

Consequently, minimising $\varepsilon_{\mathcal{B}}$ directly minimises the forgetting bound.

Role of the modal projector. Theorem 1 shows that $\widehat{P}_k$ minimises the reconstruction error $\Psi(\widehat{P}_k) = \frac{1}{ND}\sum_i \Big\|(I_T - \widehat{P}_k)\tilde{\phi}_i\Big\|_F^2$. Trajectories concentrated in the top-$\widehat{M}_k$ subspace are easier to

approximate by a small buffer: a buffer that matches the modal distribution (the $\widehat{P}_k$-projected part of $\tilde{\phi}_i$) achieves small $\varepsilon_{\mathcal{B}}$ with fewer samples. The algorithm in the main paper (Algorithm 1) selects a buffer that minimises $\varepsilon_{\mathcal{B}}$ in the whitened modal representation space, which is a computationally tractable proxy for the full trajectory approximation error (16).

Proof. **Step 1: Taylor expansion and Cauchy–Schwarz.**
Using the given update rule, a first-order Taylor expansion of $\mathcal{L}_j$ around θ_{k-1} gives:

$$\mathcal{L}_j(\theta_k) - \mathcal{L}_j(\theta_{k-1}) = -\eta \left\langle \bar{g}_j,\ \bar{g}_k - \sum_{i\in\mathcal{B}} w_i g_i \right\rangle + O(\eta^2).$$

Averaging over $j = 1, \ldots, k-1$ and applying Cauchy–Schwarz:

$$\frac{1}{k-1}\sum_{j<k}[\cdots] \ \le\ \eta\ \|\bar{g}_k\|_2\ \Big\|\frac{1}{k-1}\sum_{j<k}\bar{g}_j - \sum_{i\in\mathcal{B}} w_i g_i\Big\|_2 + O(\eta^2). \tag{18}$$

Now note that $\frac{1}{k-1}\sum_{j<k}\bar{g}_j = \frac{1}{N}\sum_{i=1}^{N} g_i$ because each $\bar{g}_j$ is an average over the same set of N samples (the past data). Denote $\bar{g} := \frac{1}{N}\sum_i g_i$. Then the gradient difference in (18) becomes $\bar{g} - \sum_{i\in\mathcal{B}} w_i g_i$.

Step 2: Affine gradient implies exact linear bound.
By Assumption 3, for each sample i, $g_i = A\,\mathrm{vec}(\phi_i) + b$ for some A and b (depending on (x_i, a_i) but fixed for the given θ_{k-1}). Hence

$$\bar{g} = A\Big(\frac{1}{N}\sum_i \mathrm{vec}(\phi_i)\Big) + b, \qquad \sum_{i\in\mathcal{B}} w_i g_i = A\Big(\sum_{i\in\mathcal{B}} w_i \mathrm{vec}(\phi_i)\Big) + b.$$

Subtracting and using the Lipschitz constant $L_\nabla = \|A\|_2$:

$$\begin{aligned}\Big\|\bar{g} - \sum_{i\in\mathcal{B}} w_i g_i\Big\|_2 &\le \|A\|_2 \Big\|\frac{1}{N}\sum_i \phi_i - \sum_{i\in\mathcal{B}} w_i\phi_i\Big\|_F \\ &= L_\nabla \Big\|\frac{1}{N}\sum_i \phi_i - \sum_{i\in\mathcal{B}} w_i\phi_i\Big\|_F. \end{aligned} \tag{19}$$

Because both averages have total weight 1, centering by $\bar{\phi}$ leaves the difference unchanged:

$$\frac{1}{N}\sum_i \phi_i - \sum_{i\in\mathcal{B}} w_i\phi_i = \frac{1}{N}\sum_i \tilde{\phi}_i - \sum_{i\in\mathcal{B}} w_i\tilde{\phi}_i.$$

Thus the right-hand side of (19) equals $L_\nabla\,\varepsilon_{\mathcal{B}}$.

Step 3: Combine. Substituting (19) into (18) yields

$$\frac{1}{k-1}\sum_{j<k}\big[\mathcal{L}_j(\theta_k) - \mathcal{L}_j(\theta_{k-1})\big] \ \le\ \eta\ \|\bar{g}_k\|_2 \cdot L_\nabla\,\varepsilon_{\mathcal{B}}\ +\ C\eta^2,$$

where C absorbs constants from the second-order remainder. □

E Theorem 2: Statistical Consistency

Theorem 2 (Finite-sample consistency)**.** *Under Assumptions 2, 4, 5, and 6, provided*

$$N\ \ge\ \frac{c_0\,T^2}{r^*\log T} \tag{20}$$

for a sufficiently large absolute constant c_0, there exist absolute constants $c_1, c_2, c_3 > 0$ such that with probability $\ge 1 - 5T^{-2}$:

$$\|\mathsf{C}_k - C_k^*\|_2\ \le\ \frac{c_1 B_\phi^2}{D}\sqrt{\frac{r^*\log T}{N}}, \tag{21}$$

$$\Big\|\widehat{P}_k - P_k^*\Big\|_F^2\ \le\ \frac{c_2\,B_\phi^4\,M^*\,r^*\log T}{\delta_{\mathrm{gap}}^2\,D^2\,N}, \tag{22}$$

$$\big|\Psi(\widehat{P}_k) - \Psi(P_k^*)\big|\ \le\ \frac{c_3\,B_\phi^2}{D}\sqrt{\frac{r^*\log T}{N}}\cdot \mathrm{tr}(\mathsf{C}_k). \tag{23}$$

Remark (Rate and sample-size condition)**.** *The dominant correction term is* $\|\frac{1}{D}\delta\delta^\top\|_2 = O(T/N)$ *(Step 3). The Bernstein term is* $O(\sqrt{r^* \log T/N})$*. The correction is dominated when* $T/N \lesssim \sqrt{r^* \log T/N}$*, i.e.* $N \gtrsim T^2/(r^* \log T)$*, which is exactly* (20)*. The rate* $O(\sqrt{r^* \log T/N})$ *is consistent with the minimax lower bound* $\Omega(\sqrt{r^*/N})$ *for operator-norm estimation of a covariance with effective rank* r^* *[Koltchinskii and Lounici, 2017].*

Proof. **Step 1: Algebraic decomposition.**
Using $\tilde{\phi}_i = \tilde{\phi}_i^* - \delta$:

$$\begin{aligned}\mathsf{C}_k &= \frac{1}{ND}\sum_i (\tilde{\phi}_i^* - \delta)(\tilde{\phi}_i^* - \delta)^\top \\ &= \frac{1}{ND}\sum_i \tilde{\phi}_i^*(\tilde{\phi}_i^*)^\top \;-\; \frac{1}{ND}\left(\sum_i \tilde{\phi}_i^*\right)\delta^\top \;-\; \frac{\delta}{ND}\left(\sum_i \tilde{\phi}_i^*\right)^\top \;+\; \frac{1}{D}\delta\delta^\top. \qquad (24)\end{aligned}$$

Since $\sum_{i=1}^N \tilde{\phi}_i^* = \sum_i(\phi_i - \bar{\phi}^*) = N\bar{\phi} - N\bar{\phi}^* = N\delta$, the two cross terms each equal $\frac{1}{D}\delta\delta^\top$, and together with the last term:

$$\mathsf{C}_k \;=\; S_N \;+\; \frac{1}{D}\delta\delta^\top \;-\; \frac{1}{D}\delta\delta^\top \;-\; \frac{1}{D}\delta\delta^\top \;+\; \frac{1}{D}\delta\delta^\top \;=\; S_N \;-\; \frac{1}{D}\delta\delta^\top, \qquad (25)$$

where $S_N = \frac{1}{N}\sum_i X_i^*$ and $X_i^* = \frac{1}{D}\tilde{\phi}_i^*(\tilde{\phi}_i^*)^\top$.

Let us verify (25) term by term. The first term in (24) is S_N. Each cross term $\frac{1}{ND}(\sum_i \tilde{\phi}_i^*)\delta^\top = \frac{1}{D}\delta\delta^\top$. The final term is $+\frac{1}{D}\delta\delta^\top$. Summing: $S_N - \frac{1}{D}\delta\delta^\top - \frac{1}{D}\delta\delta^\top + \frac{1}{D}\delta\delta^\top = S_N - \frac{1}{D}\delta\delta^\top$. Therefore:

$$\mathsf{C}_k - C_k^* \;=\; (S_N - C_k^*) \;-\; \frac{1}{D}\delta\delta^\top. \qquad (26)$$

Step 2: Bernstein bound on $S_N - C_k^*$.
The matrices $X_i^* - C_k^*$ are i.i.d., symmetric, and zero-mean. By Assumption 2, $\|X_i^*\|_2 \le R$ almost surely (Eq. (15) in the assumption), so $\|X_i^* - C_k^*\|_2 \le \|X_i^*\|_2 + \|C_k^*\|_2 \le 2R$.

Matrix variance. For any PSD matrix A with $\|A\|_2 \le R$, write $A = \sum_j \mu_j v_j v_j^\top$ with $0 \le \mu_j \le R$:

$$A^2 = \sum_j \mu_j^2 v_j v_j^\top \;\preceq\; R\sum_j \mu_j v_j v_j^\top = RA.$$

Hence $(X_i^*)^2 \preceq RX_i^*$ and:

$$V_* := \mathbb{E}\big[(X_i^* - C_k^*)^2\big] \preceq \mathbb{E}[(X_i^*)^2] \preceq R\,C_k^*.$$

Therefore $\|V_*\|_2 \le R\,\|C_k^*\|_2 = R\ \mathrm{tr}(C_k^*)/r^*$.

Applying matrix Bernstein [Tropp, 2012]. For any $t > 0$:

$$\Pr(\|S_N - C_k^*\|_2 \ge t) \le 2T\exp\left(-\frac{Nt^2/2}{\|V_*\|_2 + 2Rt/3}\right).$$

Set $t = \alpha R\sqrt{r^* \log T/N}$. Using $\|V_*\|_2 \le R\,\|C_k^*\|_2 \le R^2$ (since $\|C_k^*\|_2 \le R$):

$$\frac{Nt^2/2}{\|V_*\|_2 + 2Rt/3} \ge \frac{\alpha^2 R^2 r^* \log T/2}{R^2 + \frac{2}{3}\alpha R^2\sqrt{r^* \log T/N}} = \frac{\alpha^2 r^* \log T/2}{1 + \frac{2\alpha}{3}\sqrt{r^* \log T/N}}.$$

For $N \ge r^* \log T$, the denominator is at most $1 + \frac{2\alpha}{3}$. Choosing α so that $\frac{\alpha^2 r^*/2}{1+2\alpha/3} \ge 3$ (e.g. $\alpha = 6$ for $r^* \ge 1$), the exponent is $\le -3\log T$, giving probability $\le 2T \cdot T^{-3} = 2T^{-2}$.
Hence with probability $\ge 1 - 2T^{-2}$:

$$\|S_N - C_k^*\|_2 \;\le\; \frac{c_0' B_\phi^2}{D}\sqrt{\frac{r^* \log T}{N}}. \qquad (27)$$

Step 3: Mean-correction term.

Under Assumption 4, each entry of $\delta = \bar{\phi} - \bar{\phi}^*$ is $(\sigma/\sqrt{N})$-sub-Gaussian (average of N i.i.d. σ-sub-Gaussian variables). By a sub-Gaussian concentration bound and a union bound over the TD entries of δ, with probability $\geq 1 - 2T^{-2}$:

$$\|\delta\|_F^2 \;\leq\; \frac{c\,\sigma^2 TD}{N}.$$

Using $\|\delta\|_2 \leq \|\delta\|_F$:

$$\left\|\frac{1}{D}\delta\delta^\top\right\|_2 = \frac{\|\delta\|_2^2}{D} \leq \frac{\|\delta\|_F^2}{D} \leq \frac{c\,\sigma^2 T}{N}. \tag{28}$$

Domination condition. We need $\frac{c\sigma^2 T}{N} \leq \frac{c_0' B_\phi^2}{D}\sqrt{\frac{r^* \log T}{N}}$. Squaring and rearranging: $N \geq \frac{c^2\sigma^4 D^2 T^2}{c_0'^2 B_\phi^4 r^* \log T}$. This is exactly condition (20) with $c_0 = c^2\sigma^4 D^2/(c_0'^2 B_\phi^4)$ absorbed into the constant.

Step 4: Triangle inequality and (21)**.**

By (26) and the triangle inequality:

$$\|\mathsf{C}_k - C_k^*\|_2 \leq \|S_N - C_k^*\|_2 + \left\|\frac{1}{D}\delta\delta^\top\right\|_2 \leq \frac{c_1 B_\phi^2}{D}\sqrt{\frac{r^* \log T}{N}},$$

where the correction is absorbed into the constant c_1 under condition (20). This holds with probability $\geq 1 - 4T^{-2}$ (union bound over Steps 2 and 3).

Step 5: Davis–Kahan bound (22)**.**

Let $\varepsilon := \|\mathsf{C}_k - C_k^*\|_2$. The Davis–Kahan $\sin\Theta$ theorem states: for symmetric A, B with B having spectral gap $\delta_{\text{gap}} > 0$ at rank M^*,

$$\left\|\sin\Theta(\widehat{U}_k, U_k^*)\right\|_F \;\leq\; \frac{\sqrt{M^*}\;\|A - B\|_2}{\delta_{\text{gap}}}.$$

The standard projector identity $\|P_A - P_B\|_F^2 = 2\,\|\sin\Theta\|_F^2$ [Stewart and Sun, 1990] gives:

$$\left\|\widehat{P}_k - P_k^*\right\|_F^2 \leq \frac{2M^*\varepsilon^2}{\delta_{\text{gap}}^2} = \frac{c_2 B_\phi^4 M^* r^* \log T}{\delta_{\text{gap}}^2 D^2 N},$$

establishing (22).

Step 6: Reconstruction error (23)**.**

Write $\Psi(P) = \operatorname{tr}((I_T - P)\mathsf{C}_k)$ and split:

$$|\Psi(\widehat{P}_k) - \Psi(P_k^*)| \leq \underbrace{|\operatorname{tr}((P_k^* - \widehat{P}_k)\mathsf{C}_k)|}_{\text{(i)}} + \underbrace{|\operatorname{tr}((I_T - P_k^*)(\mathsf{C}_k - C_k^*))|}_{\text{(ii)}}.$$

Term (i). By $|\operatorname{tr}(AB)| \leq \|A\|_F\,\|B\|_F$ and $\|\mathsf{C}_k\|_F \leq \sqrt{T}\operatorname{tr}(\mathsf{C}_k)^{1/2}$ (since $\mathsf{C}_k \succeq 0$):

$$|\operatorname{tr}((P_k^* - \widehat{P}_k)\mathsf{C}_k)| \leq \left\|P_k^* - \widehat{P}_k\right\|_F \cdot \sqrt{T}\;\operatorname{tr}(\mathsf{C}_k)^{1/2} \leq \frac{\sqrt{2M^*}\,\varepsilon}{\delta_{\text{gap}}} \cdot \sqrt{T}\;\operatorname{tr}(\mathsf{C}_k)^{1/2}.$$

Term (ii). Since $\|I_T - P_k^*\|_2 = 1$: $|\operatorname{tr}((I_T - P_k^*)(\mathsf{C}_k - C_k^*))| \leq T\varepsilon$.

Both terms are bounded by $c_3\varepsilon \cdot \operatorname{tr}(\mathsf{C}_k)$ (using $\sqrt{T}\operatorname{tr}(\mathsf{C}_k)^{1/2} \leq \operatorname{tr}(\mathsf{C}_k)$ when $\operatorname{tr}(\mathsf{C}_k) \geq T$, and absorbing the constant otherwise), giving (23).

Union bound. Steps 2, 3, 4 hold with probability $\geq 1 - 2T^{-2}$, $1 - 2T^{-2}$, $1 - T^{-2}$ respectively; all hold simultaneously with probability $\geq 1 - 5T^{-2}$. □

Table 4: Train, validation, and test split counts per session and speaker for BEAT2 dataset.

Session	Speaker	Train	Val	Test
0	2	95	8	15
1	3	40	8	15
2	11	40	8	15
3	7	40	8	15
4	1	40	8	15
5	21	40	8	7
6	4	40	8	15
7	6	40	10	9
8	9	40	8	7

Table 5: Train, validation, and test split counts per session and speaker for 3D mesh dataset in TALKSHOW.

Session	Name (Speaker)	Train	Val	Test
0	Oliver	96	12	13
1	Seth	44	15	15
2	Chemistry	55	32	32

F Dataset Details

We evaluate TMR on three benchmarks spanning distinct audio modalities, motion representations, and continual distribution shift axes. Each benchmark is organized as a sequence of discrete sessions, where session 0 serves as the base training step and all subsequent sessions are introduced sequentially without access to prior raw data. Below we describe the data source, preprocessing decisions, session ordering rationale, and split construction for each benchmark.

F.1 BEAT2 with GestureLSM

The BEAT2 dataset [Liu et al., 2024] is a large-scale, motion-captured co-speech gesture corpus containing over 60 hours of SMPL-based full-body motion synchronized with speech audio from multiple speakers. For our continual speaker learning benchmark, we select nine speakers: Wayne, Scott, Solomon, Lawrence, Carla, Sophie, Miranda, Nidal, and Ayana. Speaker selection was guided by two criteria: (i) substantial variation in body morphology and movement style across the subset, and (ii) sufficient per-speaker data to enable stable training within each session. This produces nine sequential sessions, each introducing one new speaker, forming the most extensive continual speaker-learning benchmark in this work.

The session ordering follows the ascending speaker-ID sequence (2, 3, 11, 7, 1, 21, 4, 6, 9), which is fixed prior to any hyperparameter tuning and is not chosen to favor any specific method. Each session uses a train/validation/test as given in Table 4. Session 0 (Speaker 2) is allocated the largest training pool (95 clips) to provide a stable initialization; all subsequent sessions use 40 training clips, 8 validation clips, and between 7 and 15 test clips depending on speaker availability, as summarized in Table 4. All clips are pre-processed to a fixed temporal length of T frames by uniform trimming or zero-padding, with audio features extracted using the WavLM encoder following the GestureLSM preprocessing pipeline [Liu et al., 2025].

The nine-session structure makes BEAT2 the most challenging benchmark in terms of long-horizon forgetting pressure: by Session 8, the replay buffer must simultaneously represent eight prior speakers across substantially different gesture vocabularies, providing a rigorous stress test for any continual learning method.

F.2 TalkSHOW with VQ-VAE

The TalkSHOW dataset [Yi et al., 2023] consists of holistic 3D human mesh sequences derived from 26.9 hours of in-the-wild video footage, with SMPL-X parameters reconstructed via pseudo-ground-truth fitting. The dataset covers three distinct speakers: Oliver, Seth, and Chemistry, each characterized by a unique speaking style, gesture vocabulary, and motion dynamic range. Unlike BEAT2, which provides motion-captured ground truth, TalkSHOW data is obtained through video-

Table 6: Train, validation, and test split counts per genre for SAM dataset in MOSPA.

Session	Genre	Train	Val	Test
0	Insensitive	818	106	96
1	Sensitive	829	88	98
2	Neutral	822	100	98

based reconstruction, introducing a higher degree of per-speaker motion variability and reconstruction noise that is itself speaker-dependent—making forgetting more challenging to mitigate.

The continual learning curriculum proceeds over three sessions in the order Oliver $\rightarrow$ Seth $\rightarrow$ Chemistry, reflecting the chronological speaker order in the original benchmark. Session 0 initializes the VQ-VAE generator on Oliver (96 training, 12 validation, 13 test clips), Session 1 introduces Seth (44/15/15), and Session 2 introduces Chemistry (55/32/32), as reported in Table 5. The larger validation and test allocations for Chemistry and Seth reflect the greater total clip availability for those speakers after training split reservation. Note that the Chemistry split uses substantially more validation and test clips than Oliver, providing a more robust final evaluation of the model's generalization to the most recently introduced speaker.

The VQ-VAE architecture of TalkSHOW encodes body and hand motion into separate discrete codebooks, making continual fine-tuning particularly susceptible to codebook drift: gradient updates from a new speaker can overwrite gesture-token distributions learned for prior speakers, causing the decoder to hallucinate incorrect tokens when conditioned on old-speaker audio. This makes TalkSHOW a qualitatively distinct forgetting challenge from the flow-matching setting of GestureLSM and an important testbed for the generalization of TMR across generator families.

F.3 SAM with MOSPA

The SAM dataset [Xu et al., 2025] contains over 9 hours of Vicon-captured full-body SMPL-X motion recorded under 27 distinct spatial audio scenarios and 49 action categories, with reactions spanning three qualitatively different motion genres that characterize the degree of bodily responsiveness to acoustic stimuli. Each sample in the dataset carries an integer genre label $g \in \{0, 1, 2\}$ encoded in its identifier:

- **Genre 0 — Insensitive.** The character exhibits minimal or no reaction to surrounding audio. Motions are dominated by idle standing, postural sway, and subtle weight shifts, with negligible audio-motion correlation. This genre defines the lowest-variance, lowest-energy regime in the dataset.
- **Genre 1 — Neutral.** The character displays moderate reactive behavior, including head turns toward sound sources, occasional limb adjustments, and casual postural repositioning, reflecting an intermediate level of acoustic awareness and audio-motion coupling.
- **Genre 2 — Sensitive.** The character produces strong, full-body reactive motion: covering the ears, flinching, turning away, or actively moving in response to the spatial audio stimulus. These motions exhibit the highest audio-motion coupling, greatest kinematic diversity, and most pronounced temporal variation.

We organize the SAM continual learning curriculum into three sessions with the deliberate ordering Insensitive $\rightarrow$ Sensitive $\rightarrow$ Neutral, as detailed in Table 6. This ordering is chosen specifically to maximize distributional contrast between consecutive sessions: the transition from Session 0 to Session 1 presents the largest possible shift in motion distribution (from the lowest-energy idle regime to the highest-energy reactive regime), creating maximum forgetting pressure on any method that lacks explicit retention mechanisms. Session 2 then introduces the Neutral genre, whose intermediate distribution lies between the two extremes, testing whether the model can simultaneously preserve knowledge of both prior genres while acquiring moderate-reactivity motion.

Each genre provides approximately 820 training, 95 validation, and 97 test clips (Table 6), with splits defined by stratified sampling across spatial audio scenario and action category to ensure that no single scenario or action type dominates any partition. Audio features are binaural spatial audio embeddings extracted following the MOSPA preprocessing pipeline [Xu et al., 2025], including sound source azimuth and elevation encoded in the character's local frame. Unlike the speaker-adaptation settings of BEAT2 and TalkSHOW, the SAM benchmark shifts the continual axis from identity to behavioral mode, making it the only benchmark in this work where the input audio modality (spatial

audio rather than speech) and the nature of the distribution shift (motion genre rather than speaker style) differ fundamentally from the other two—providing a comprehensive and complementary test of TMR's architecture-agnostic generalization.

G Additional Results

Table 7: **Replay construction complexity on GestureLSM (BEAT2) at Session 4.** TMR achieves substantially lower forgetting with replay construction cost and VRAM comparable to existing replay methods, while introducing no additional inference overhead or model parameters.

Method	Training Time (mins)	Buffer Creation Time (min:sec)	Buffer VRAM (MiB)
OCS	570.41	1:27	2970
TMR	573.22	1:46	2028
MIR	574.58	2:11	1336
GCR	567.80	3:31	1996
CSReL	569.21	4:29	5350

G.1 Replay Construction Complexity

Table 7 reports replay buffer construction time, training time, VRAM consumption for buffer creation for TMR and competing replay baselines on GestureLSM (BEAT2) at Session 4, where the replay pool spans four prior speakers. TMR constructs its buffer in 1:46 minutes, substantially faster than GCR (3:31 minutes) and CSReL (4:29 minutes), while consuming only 2028 MiB of VRAM during buffer construction — lower than OCS (2970 MiB) and less than half that of CSReL (5350 MiB). Training times are tightly clustered across all methods (567–575 minutes), confirming that the overhead of TMR's eigendecomposition, rank selection, and greedy sparse recovery is negligible relative to the cost of training the generator itself. Crucially, once buffer construction is complete, all methods train under identical conditions with the same replay size, model capacity, and number of iterations, meaning that TMR's substantially lower forgetting is attributable entirely to the quality of its replay selection rather than any additional computational advantage at training time. TMR introduces zero additional trainable parameters and imposes no overhead at inference.

G.2 Long-Horizon Continual Learning on GestureLSM

Table 8 extends the GestureLSM evaluation to the full nine-session continual speaker-learning benchmark on BEAT2, comparing vanilla fine-tuning against TMR across all speakers and sessions. The results reveal a striking and consistent pattern: vanilla forgetting compounds monotonically with session count, with Speaker 2 degrading from an initial FGD of 0.412 at Session 0 to 1.900 by Session 8 ($+361\%$), while Speakers 7 and 6 reach FGD values of 5.548 and 5.053 respectively by the final session. In contrast, TMR maintains Speaker 2 near-flat across all nine sessions, ending at $\mathbf{0.601}$ — remarkably close to the Session 0 baseline — while consistently outperforming vanilla on every prior speaker at every session checkpoint. The aggregate forgetting metric F^{avg} confirms this: TMR achieves $\mathbf{0.379}$ versus 1.384 for vanilla, a reduction of over 72%, while maintaining virtually identical adaptation performance (A^{avg} of 2.728 vs. 2.714). Importantly, TMR's per-speaker FGD values fluctuate only minimally between Sessions 4 and 8, demonstrating that forgetting does not compound even as the replay buffer must represent an increasingly heterogeneous set of speakers. This directly validates the theoretical prediction of Section 3.3: as sessions accumulate, the temporal feature operator C_k captures progressively richer cross-speaker covariance structure, making the modal replay representation more representative over time rather than less.

H Why Continual Learning Baselines Fail in Audio-Conditioned 3D Motion Generation

Standard replay and regularization methods were designed for image classification, not for the unique challenges of audio–driven motion. TMR succeeds where they fail because it preserves the temporal structure of audio–feature trajectories. The table below catalogues the fatal flaw of each baseline.

H.1 Detailed Failure Analysis

1. **Coreset via Reducible Loss (CSReL)** Tong et al. [2025]: CSReL selects samples that maximally reduce loss when added to training. In audio–motion, ambiguous or noisy samples have high loss even on full data, so they are wrongly prioritised. The modal whitening in TMR filters such noise by projecting onto high-variance temporal modes, retaining only those trajectories that align with the dominant audio–motion correlations.
2. **Generative Replay (GenRep)** Van de Ven and Tolias [2018]: GenRep uses a VAE to generate pseudo-samples of past data. For audio–motion, generated feature trajectories often lose frame-wise rhythmic alignment, accumulating drift over sessions. TMR replays *real* stored trajectories, guaranteeing exact audio–time alignment without generative noise.
3. **C-Flat + replay** Bian et al. [2024]: C-Flat flattens the loss landscape in parameter space via sharpness-aware minimisation. This makes the model robust to small weight changes but does not protect task-specific *temporal modes*. When new tasks arrive, previously learned temporal correlations are overwritten. TMR explicitly maintains those modes via the modal projector.
4. **STAR + replay** Eskandar et al. [2025]: STAR introduces a worst-case perturbation of model parameters to stabilise predictions. However, the perturbation is isotropic in the weight space, ignoring the fact that forgetting is caused by misalignment of *trajectory-level* features. TMR's trajectory-level error bound directly controls forgetting.
5. **UCL (Adaptive Regularization)** Ahn et al. [2019]: UCL estimates per-weight uncertainties to penalise changes. This is effective when tasks differ in static feature importance, but fails when tasks differ in *temporal dynamics*. Two tasks may use the same weight for different timing patterns; UCL cannot disambiguate them. TMR separates the temporal axis via eigendecomposition of C_k.
6. **Online Coreset Selection (OCS)** Yoon et al. [2021]: OCS selects samples greedily by gradient alignment with the current task. In a streaming audio environment, a single sample may have high gradient alignment due to a few noisy frames, causing it to be selected even if its overall trajectory is irrelevant. TMR's whitened representation normalises per-mode energy, preventing any single time frame from dominating selection.
7. **Maximally Interfered Retrieval (MIR)** Aljundi et al. [2019]: MIR retrieves samples whose predictions would change most after a parameter update. This criterion is highly sensitive to label noise and assumes i.i.d. data. In audio–motion, small perturbations in frame alignment can cause large changes in inference, leading to retrieval of uninformative samples. TMR's structured replay via the modal subspace is robust to such local misalignments.
8. **Pareto-Optimized CL (POCL)** Wu et al. [2024]: POCL casts CL as multi-objective optimisation and seeks a Pareto-optimal gradient. When audio tasks have orthogonal temporal signatures (e.g., rhythmic gestures vs. continuous motion), any Pareto front collapses to a compromise that retains neither. TMR's modal decomposition separately represents each temporal mode, allowing the buffer to cover the union of task-specific subspaces.
9. **Gradient Coreset Replay (GCR)** Tiwari et al. [2022]: GCR selects a coreset that approximates the full gradient in parameter space. The gradient of a $T \times D$ trajectory is compressed into a single vector, discarding all temporal structure. TMR's trajectory-level error $\varepsilon_{\mathcal{B}}$ keeps the full $T \times D$ shape and bounds gradient mismatch via the affine property.
10. **Experience Replay (ER)** Rolnick et al. [2019]: Uniform random sampling from a buffer treats each snapshot as independent, destroying the sequential correlation between frames. For audio–motion, adjacent frames contain crucial timing information; random sampling breaks this. TMR's modal replay preserves the joint distribution of whole trajectories.
11. **Elastic Weight Consolidation (EWC)** Kirkpatrick et al. [2017]: EWC adds a quadratic penalty on parameter deviations weighted by Fisher information. This penalty treats all time steps equally and cannot differentiate between parameters that influence early vs. late frames. TMR's temporal projector $\widehat{P}_k$ applies a time-selective mask, protecting only the modes relevant to past tasks.
12. **SD-LoRA** Wu et al. [2025]: SD-LoRA decouples magnitude and direction of low-rank adaptation matrices, enabling parameter-efficient CL without replay. However, it works on

static pre-trained features and ignores sequence-level dynamics. For audio–motion, the same LoRA weight may be reused for different temporal patterns, causing overwriting. TMR's explicit buffer of past trajectories avoids any parameter-level interference.

In summary, all 12 baselines share a common blind spot: they operate on gradients, weights, or sample-wise scores, discarding the temporal structure of audio–feature trajectories. TMR succeeds because it preserves that structure via the temporal feature operator C_k and the whitened modal representation γ_i, aligning perfectly with the theoretical guarantees.

I Limitations and Future Work

Despite strong continual retention across flow-matching, autoregressive, and diffusion-based generators, TMR still relies on a bounded replay buffer whose memory grows with session count as with other replay based methods. Developing theoretically grounded *memory compression*, hierarchical replay, or *streaming eigenspace approximations for extremely long continual trajectories* remains an important direction. Our formulation currently models temporal covariance structure within audio-conditioned trajectories only; **extending TMR to jointly model cross-modal geometry between audio, text, facial expression, gaze, and full-body motion may further improve robustness under complex multimodal shifts**. Another promising direction is replacing static replay selection with *adaptive online temporal subspaces* that evolve continuously under streaming data without explicit session boundaries. More broadly, the temporal feature operator introduced in this work may provide a foundation for continual learning beyond motion generation, including video diffusion, embodied agents, humanoid robotics, and multimodal world models, where preserving long-horizon temporal structure is fundamentally important. Finally, integrating temporal-modal replay with parameter-efficient adaptation, generative replay, or foundation-scale motion models may enable scalable lifelong human behavior modeling over open-world environments.

Table 8: FGD scores for Vanilla and TMR across 9 sessions of BEAT2 speakers. Lower is better.

Session	Speaker	Tested On	Vanilla	TMR
0	2	—	0.412	
1	3	2	1.489	0.454
		3	2.260	2.229
2	11	2	2.140	0.488
		3	3.160	2.483
		11	2.059	2.078
3	7	2	1.874	0.604
		3	3.241	2.551
		11	2.964	2.926
		7	3.508	3.571
4	1	2	2.129	0.613
		3	3.215	2.605
		11	2.972	2.624
		7	4.112	3.570
		1	2.392	2.565
5	21	2	1.797	0.666
		3	3.085	2.524
		11	3.052	2.665
		7	4.549	3.758
		1	3.313	2.734
		21	4.031	3.993
6	4	2	1.362	0.600
		3	2.714	2.312
		11	2.873	2.613
		7	4.105	3.443
		1	3.166	2.690
		21	5.119	4.551
		4	1.966	1.969
7	6	2	2.301	0.740
		3	3.313	2.560
		11	3.199	2.737
		7	4.520	3.623
		1	3.941	2.925
		21	5.291	4.641
		4	3.469	2.309
		6	2.867	2.884
8	9	2	1.900	0.601
		3	3.250	2.450
		11	3.688	2.772
		7	5.548	3.754
		1	4.270	2.908
		21	6.257	5.018
		4	3.952	2.507
		6	5.053	3.331
		9	2.626	2.538
		$F^{avg} \downarrow$	1.384	**0.379**
		$A^{avg} \downarrow$	**2.714**	2.728